CLAIMS

What is claimed is:

1. A method for balancing a load in a network having a load balancing slave, a load balancing master, a plurality of servers, and a client, the method comprising the steps of:

receiving at the load balancing slave a request from the client to perform processing;

sending by the load balancing slave the request to the load balancing master in response to the receipt of the request;

determining a load of each of the plurality of servers by the load balancing master;

selecting by the load balancing master a selected one of the plurality of servers that is suitable for performing the processing, wherein the selected server is selected based on the load of each of the plurality of servers;

sending an identifier of the selected server from the load balancing master to the load balancing slave; and

establishing by the load balancing slave a communication link between the selected server and the client to perform the processing.

2. The method of claim 1, wherein the step of establishing further includes the step of:

routing the communication link between the selected server and the client through the load balancing slave.

3. The method of claim 1, further comprising the step of:

receiving a plurality of load metrics from each of the plurality of servers.

4. The method of claim 1, wherein the step of determining further comprises the step of:

receiving a load metric with the request from the load balancing slave at the load balancing master.

5. A method in a data processing system having a first and a second load balancing server and having a plurality of processing servers, the method comprising the steps of:

receiving by the first load balancing server a request to perform processing;

sending the request from the first load balancing server to the second load balancing server;

determining a load of each of the processing servers by the second load balancing server;

selecting by the second load balancing server a selected one of the plurality of processing servers that is suitable for performing the processing , wherein the selection is performed based on the load of each of the plurality of processing servers; and

sending by the second load balancing server to the selected processing server an indication to perform the processing.

6. The method of claim 5, wherein the step of sending by the second load balancing server further comprises the step of:

identifying to the first load balancing server the selected server after the indication to perform the processing has been sent to the selected processing server.

7. The method of claim 5, further comprising the steps of:

receiving a plurality of load metrics that originate from the plurality of processing servers at the second load balancing server.

8. The method of claim 5, wherein sending a request further includes the step of;

encoding the at least one load metric in the request.

9. The method of claim 5, wherein the first load balancing processor is a load balancing slave.

10. The method of claim 5, wherein the second load balancing processor is a load balancing master.

11. A data processing system, comprising:

a plurality of processing servers;

a client sends a request;

a load balancing slave that receives the request from the client, that sends the request to an external source for a selection of one of the plurality of processing servers that is suitable for performing the processing, that receives an indication of the selected processing server from the external source, and that establishes a communication link between the selected processing server and the client to perform the processing; and

a load balancing master that receives the request from the load balancing slave, that determines a load of each of the plurality of processing servers, that selects the selected

processing server based on the load of each of the plurality of processing servers, and that sends the indication of the selected server to the load balancing slave.

12. The data processing system of claim 11, wherein a plurality of load metrics are received at the load balancing master from the plurality of processing servers that indicate the load on each of the plurality of processing servers.

13. The data processing system of claim 11, wherein at least one load metric is included in the request sent by the load balancing slave to the external source.

14. A data processing system, comprising:

a plurality of processing servers;

a client that sends a request to have processing performed in a load balanced manner;

a first load balancing server that receives the request from the client; and

a second load balancing server that receives the request from the first load balancing server, that determines a load of each of the processing servers, that selects a selected one of the plurality of processing servers that is suitable for performing the processing in the load balanced manner, and that sends to the selected processing server an indication to perform the processing, wherein the selection is based on the load of each of the plurality of processing servers.

15. The data processing system of claim 14, wherein the first load balancing server is a load balancing slave.

16. The data processing system of claim 14, wherein the second load balancing server is a load balancing master.

17. The data processing system of claim 14, wherein the second load balancing server is in receipt of a plurality of load metrics that originating from each of the processing servers and indicate the load on each of the processing servers.

18. A computer-readable medium containing instructions that cause a data processing system to perform a method for balancing a load in a network having a load balancing slave, a load balancing master, a plurality of servers, and a client, the method comprising the steps of:

receiving at the load balancing slave a request from the client to perform processing;

sending by the load balancing slave the request to the load balancing master in response to the receipt of the request;

determining a load of each of the plurality of servers by the load balancing master;

selecting by the load balancing master a selected one of the plurality of servers that is suitable for performing the processing, wherein the selected server is selected based on the load of each of the plurality of servers;

sending an identifier of the selected server from the load balancing master to the load balancing slave; and

establishing by the load balancing slave a communication link between the selected server and the client to perform the processing.

19. The computer-readable medium of claim 18, wherein the step of establishing further includes the step of:

routing the communication link between the selected server and the client through the load balancing slave.

20. The computer readable medium of claim 18, further comprising the step of:

receiving a plurality of load metrics from each of the plurality of servers.

21. The computer readable medium of claim 18, wherein the step of determining further comprises the step of:

receiving a load metric with the request from the load balancing slave at the load balancing master.

22. A computer readable medium containing instructions that cause a data processing system to perform a method for load balancing having a first and a second load balancing server and having a plurality of processing servers, the method comprising the steps of:

receiving by the first load balancing server a request to perform processing;

sending the request from the first load balancing server to the second load balancing server;

determining a load of each of the processing servers by the second load balancing server;

selecting by the second load balancing server a selected one of the plurality of processing servers that is suitable for performing the processing, wherein the selection is performed based on the load of each of the plurality of processing servers; and

sending by the second load balancing server to the selected processing server an indication to perform the processing.

23. The method of claim 22, wherein the step of sending by the second load balancing server further comprises the step of:

identifying to the first load balancing server the selected server after the indication to perform the processing has been sent to the selected processing server.

24. The computer-readable medium of claim 22, further comprising the steps of:

receiving a plurality of load metrics that originate from the plurality of processing servers at the second load balancing server.

25. The computer-readable medium of claim 22, wherein sending a request further includes the step of;

encoding the at least one load metric in the request.

26. A load balancer for balancing a load in a network having a load balancing slave, a load balancing master, a plurality of servers, and a client, the method comprising the steps of:

means for receiving at the load balancing slave a request from the client to perform processing;

means sending by the load balancing slave the request to the load balancing master in response to the receipt of the request;

means for determining a load of each of the plurality of servers by the load balancing master;

means for selecting by the load balancing master a selected one of the plurality of servers that is suitable for performing the processing, wherein the selected server is selected based on the load of each of the plurality of servers;

means for sending an identifier of the selected server from the load balancing master to the load balancing slave; and

means for establishing by the load balancing slave a communication link between the selected server and the client to perform the processing.